Classification: Naive Bayes Classifier

by shal21

For this assignment, we will be using the `Zoo Animal Classification` dataset from the UCI Machine Learning dataset. https://archive.ics.uci.edu/ml/datasets/Zoo

This dataset consists of 101 animals from a zoo. There are 16 variables with various traits to describe the animals. The 7 Class Types are: Mammal, Bird, Reptile, Fish, Amphibian, Bug and Invertebrate.

For grading purposes, we will be using subsets of this public dataset, so please train your model on our provided data.

Before you start to work on the assignment, let's review what we learned from class:

Naive Bayes Classifier

Let's denote the features as $\boldsymbol{X}$ and the label as $\boldsymbol{y}$. As a generative model, the naive Bayes classifier makes predictions based an estimation of the joint probability $\boldsymbol{P(X, y)}$.

For each example, the predicted label $\hat{\boldsymbol{y}}$ is determined by:

$$\hat{y} = argmax_y P(X, y) = argmax_y P(y)P(X|y)$$

In the naive Bayes classifier, we make the assumption that all features are independent given the class label. This means that:

$$P(X|y) = \prod_i P(x_i|y)$$

.

To make our prediction, we need to keep track of $\boldsymbol{P(y)}$ and $\boldsymbol{P(x_i|y)}$.

Smoothing

Since some data combinations do not appear in our dataset, we smooth out the probability $\boldsymbol{P(x_i|y)}$ and $\boldsymbol{P(y)}$ with Laplacian correction. Specifically, since our dataset is small, we smooth the probability with a psuedo-count of 0.1.

That is,

$$P(y = c) = \frac{\#\text{ of samples that have class c} + 0.1}{|N| + 0.1|C|}$$

$$P(x_i = f|y = c) = \frac{\#\text{ of samples that have class c}, x_i \text{is f} + 0.1}{\#\text{ of samples that have class c} + 0.1\#\text{ of unique features f}}$$

Input Format

CSV file with the following fields: animal_name, hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes,

venomous, fins, legs, tail, domestic, catsize, classtype. Do not use the `animal_name` and `classtype` for prediction.

The examples will be split into training and test set. For the training set, the `classtype` field will have an integer value from 1-7. For the test set, the `classtype` field will have the value -1.

- animal_name: Unique for each instance
- hair Boolean
- feathers Boolean
- eggs Boolean
- milk Boolean
- airborne Boolean
- aquatic Boolean
- predator Boolean
- toothed Boolean
- backbone Boolean
- breathes Boolean
- venomous Boolean
- fins Boolean
- legs Numeric (set of values: {0,2,4,5,6,8})
- tail Boolean
- domestic Boolean
- catsize Boolean
- class_type Numeric (integer values in range [1,7])

Constraints

You are not allowed to use directly use machine learning libraries such as `sklearn`.

Output Format

For each example in the test set, print the predicted `classtype`.

Sample Input 0

```
animal_name,hair,feathers,eggs,milk,airborne,aquatic,predator,toothed,backbone,breathes,venomous,fins,legs,tail,d
omestic,catsize,class_type
aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1
worm,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,7
piranha,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4
gnat,0,0,1,0,1,0,0,0,0,1,0,0,6,0,0,0,6
oryx,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
moth,1,0,1,0,1,0,0,0,0,0,1,0,0,6,0,0,0,6
skimmer,0,1,1,0,1,1,1,0,1,1,0,0,2,1,0,0,2
crab,0,0,1,0,0,1,1,0,0,0,0,0,4,0,0,0,7
vampire,1,0,0,1,1,0,0,1,1,1,0,0,2,1,0,0,1
slowworm,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,3
bass,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,-1
```

Sample Output 0

```
4
```

Explanation 0

The joint probability of the predicted class is 0.01916.

Submissions: 158
Max Score: 5
Difficulty: Medium

Rate This Challenge:

More

Current Buffer (saved locally, editable) Python 3

```
1 import sys
2 import csv
3 zoodataset=[]
4 for row in csv.reader(iter(sys.stdin.readline, '')):
5     zoodataset=zoodataset+row
6 print(zoodataset)
```

Line: 1 Col: 1

Upload Code as File Test against custom input Run Code Submit Code